Proceedings of the 10th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'24)
Barcelona, Spain - August 19 - 21, 2024
Paper No. CIST 179
DOI: 10.11159/cist24.179

Prediction of Severity Level of Road Traffic Accident in Thailand using Machine Learning

Supitcha Ratanawimon, Patrawadee Tanawongsuwan[1]
Graduate School of Applied Statistics, National Institute of Development Administration
148 Serithai Road, Bangkapi, Bangkok Thailand
[1]patrawadee@as.nida.ac.th

Abstract **-** Road traffic accidents have been increasing rapidly worldwide due to the continuous growth in population and vehicles. The purpose of this research was to use Machine Learning algorithms to create a model for predicting the severity of road traffic accidents. The research focused on the Extreme Gradient Boosting (XGBoost) algorithm for prediction and compared its performance with four other algorithms, namely Random Forest, Bagging, Decision Tree, and Multilayer Perceptron. The research methodology covered several essential steps, including data preprocessing, class weighting, model building, and performance evaluating using appropriate metrics. The results revealed that the XGBoost model outperformed the other models in predicting the severity of road traffic accidents, especially accidents of fatal severity. The model achieved precision of 78%, recall of 57%, F1-score of 66%, balanced accuracy of 77%, and an impressive ROC-AUC of 90%. The results could be utilized in strategic planning and implementing appropriate measures to reduce and prevent road traffic accidents in Thailand.

Keywords**:** Machine Learning, Extreme Gradient Boosting, Severity level prediction of road traffic accidents, Road traffic accident

1. Introduction

The number of road accidents has been increasing rapidly worldwide due to the continuous rise in population and vehicles. According to data from the World Health Organization (WHO) [1], approximately 1.3 million people die from road traffic accidents every year, and 20-50 million people are injured. This has led to significant losses and impacts for many countries. Therefore, it is necessary to implement road safety action plans to continually reduce and prevent road traffic accidents. Accurate prediction of road accident severity plays a crucial role in developing effective strategies for reducing and preventing accidents.

This research focused on creating a model to predict the severity of road traffic accidents in Thailand on different types of roads using supervised machine learning algorithms. The severity of road traffic accidents was classified into two levels: Fatal and Injured. The data analyzed in this study comprised 81,735 accident records from the road network of the Ministry of Transport [2], collected from the Ministry of Transport of Thailand between 2019 and 2022. The dataset included various attributes related to accident location, road configuration, suspected cause, accident types, weather conditions, and more.

For the machine learning technique, the researcher selected the Extreme Gradient Boosting (XGBoost) algorithm to create a model for predicting the severity of road traffic accidents. This algorithm was chosen due to its high efficiency and because it had not been widely used in previous research. Additionally, the researcher selected various algorithms to compare the model's performance, including Random Forest (RF), Bootstrap Aggregation (Bagging), Decision Tree (DT), and Multilayer Perceptron (MLP). The evaluation metrics used to assess the performance include Precision, Recall, F1-Score, Balanced Accuracy, and the ROC-AUC curve (Receiver Operating Characteristic - Area Under the Curve) analysis.

The objectives of this study were as follows: (1) to create a model for predicting the severity of road traffic accidents using various machine learning algorithms; (2) to evaluate and compare the performance of the models using evaluation metrics; (3) to analyze the data for insights into the factors causing road traffic accidents that resulted in injuries or fatalities, and to present the research findings in the most beneficial way to support the development of strategies for reducing and preventing accidents.

2. Literature Review

From the review of literature on related studies, it was found that Hmamed Hala et al. [3] developed a model for predicting the severity of road traffic accidents using real accident data from New Zealand. They employed various machine learning algorithms to classify accident severity into three levels: Fatal, Serious, and Slight. They compared the performance of different models, and the results indicated that the top three models with the highest accuracy, precision, recall, and ROC-AUC values were Multilayer Perceptron, Support Vector Machine, and K-Nearest Neighbors, respectively.

In the research by Sumbal Malika et al. [4], a model was developed to predict the severity of road traffic accidents using real accident data from the United Kingdom. This study utilized the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance in the dataset and employed machine learning algorithms to classify accident severity into three levels: Fatal, Serious, and Slight. The results showed that the top three models with the highest accuracy, precision, and recall were Random Forest, Decision Tree, and Bagging, respectively.

In this study, the researcher chose an algorithm of particular interest that had not been explored in previous studies, specifically **Extreme Gradient Boosting (XGBoost)** [5], to further study and predict the severity of road traffic accidents. XGBoost is an efficient model developed from gradient boosting. It is a distributed gradient boosting library optimized for efficient and scalable training of machine learning models. It is an ensemble learning method that combines the predictions of multiple weak models to create a stronger prediction. One of the key capabilities of XGBoost is its ability to address overfitting through regularization, a technique used to reduce the variance of machine learning models. Additionally, it can perform parallel processing for large datasets, making it highly suitable for handling big data in predicting road traffic accident severity.

Additionally, the researcher selected four machine learning algorithms from the previously mentioned studies, which demonstrated high overall predictive performance, to compare their effectiveness against the Extreme Gradient Boosting algorithm. These algorithms are outlined in the paragraphs that follow.

Random Forest (RF) [6] is part of a group of models known as ensemble learning methods, specifically multiple-learner techniques. It consists of a large number of individual decision trees that operate as a collective. Each tree in the random forest makes a prediction for the class, and the class that receives the most votes from all the trees is selected as the final prediction.

Bootstrap Aggregation (Bagging) [7] is another ensemble learning method that falls under the category of multiple-learner techniques. It involves creating multiple subsets of the original dataset through random sampling with replacement. The principle behind Bagging is to create several bootstrapped samples, where each bootstrapped sample is typically two-thirds the size of the original dataset. Since the sampling is done with replacement, it is possible for the same data point to appear multiple times within a single bootstrapped sample.

Decision Tree (DT) [8] is a hierarchical structure that resembles a flowchart, where each node represents an attribute, and each branch represents a decision based on that attribute. The leaf nodes represent the outcome or prediction. The algorithm learns to make decisions by splitting the data into subsets based on different attributes, selecting the attribute that best separates the data into different classes or reduces uncertainty.

Multilayer Perceptron (MLP) [9] is a type of Artificial Neural Network (ANN) that consists of nodes or neurons connected in multiple layers. It is a feedforward neural network, meaning that data flows from the input layer through hidden layers to the output layer. MLP is capable of learning complex patterns and predicting based on the relationships learned between inputs and outputs.

3. Research Methodology

The objective of this research was to create a model to predict the severity of road traffic accidents. The proposed research process, illustrated in Fig. 1, consisted of the following steps:

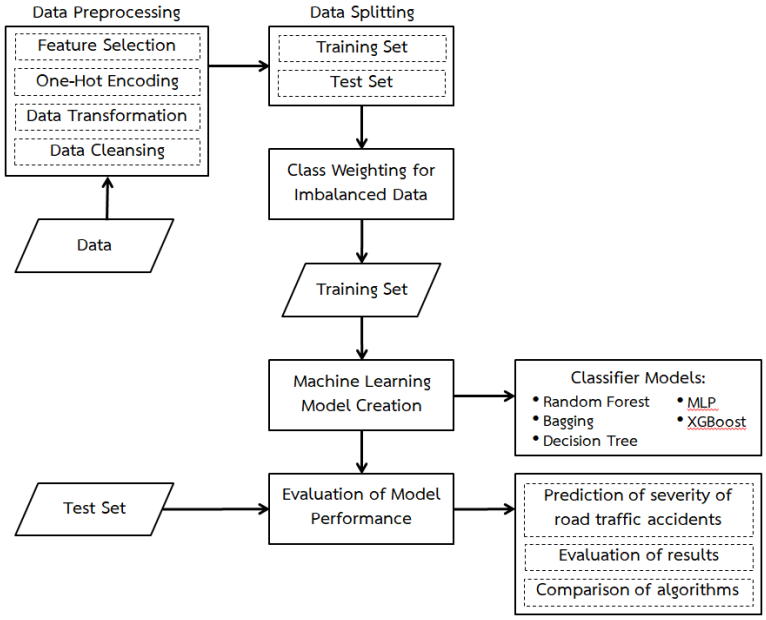


Fig. 1: Resesarch process

3.1. Datasets

This research obtained the dataset from the website of the Ministry of Transport [2], which consists of road traffic accidents that occurred on national highways, rural roads, and expressways in Thailand from the year 2019 to 2022, totaling 81,735 accidents. The dataset comprises 20 attributes as shown in Table 1, with the severity of road traffic accidents categorized into 2 levels: Fatal and Injured.

Table 1: The attributes of the road accident dataset from the Ministry of Transport's road network

No.	Attributes	No.	Attributes
1	Year	11	Vehicle1_Type (such as motorcycle, car, pick-up truck, etc.)
2	Date	12	Road_Configuration (such as intersection, straight, sharp curve, etc.)
3	Time	13	Suspected_Cause (such as drunk driving, slippery road, sudden stop, etc.)
4	Report_Date	14	Accident_Type (such as rear-end collision, turning/reversing collision, overturn, etc.)
5	Report_Time	15	Number_of_Vehicles
6	Case_Number	16	Number_of_Deaths
7	Road_Type (highway, rural road, or expressway)	17	Number_of_Injuries
8	Accident_Location (such as route number, intersection name, location name, etc.)	18	Weather_Condition (such as clear, raining, foggy, etc.)
9	Kilometer_Marker	19	Latitude
10	Province	20	Longitude

3.2. Data Preprocessing

Data preprocessing is the process of preparing data for machine learning by ensuring its consistency, completeness, and accuracy in the learning environment. It can be carried out through the following steps:

1.Data Cleansing: Removing unnecessary attributes and handling missing values.
2.Data Transformation: Reformatting data for analysis and modeling.
3.One Hot Encoding: Converting categorical variables to binary columns.
4.Feature Selection: Using the Chi-square test to find the most important attributes. A high value would suggest a strong relationship with the target variable. The top ten attributes based on Chi-square values are shown in Table 2.

Table 2: The top ten attributes by Chi-square values

Attribute	Chi-square Values
Vehicle1_Type=motorcycle	3706.14
Accident_Type=pedestrian_collision	1729.78
Accident_Type=opposite_direction_collision_not_overtaking	610.50
Vehicle1_Type=car_private_or_public	274.36
Accident_Type=rollover_or_off_road_in_straight_path	272.46
Suspected_Cause=sudden_obstruction_by_person_vehicle_animal	157.16
Province=bangkok	154.38
Vehicle1_Type=truck_4_wheel_pickup	150.48
Accident_Type= collision_with_obstacle_on_road_surface	122.41
Province=chonburi	89.85

Fig. 2-5 show the top three Chi-square values for select attributes that have demonstrated high values.

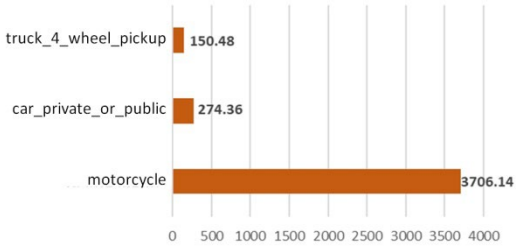


Fig. 2: Top 3 Chi-square values for Vehicle1_Type

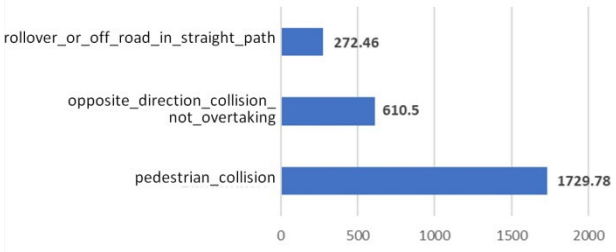


Fig. 3: Top 3Chi-square values for Accident_Type

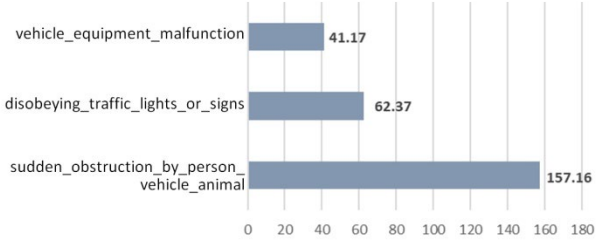


Fig. 4: Top 3 Chi-square values for Suspected_Cause

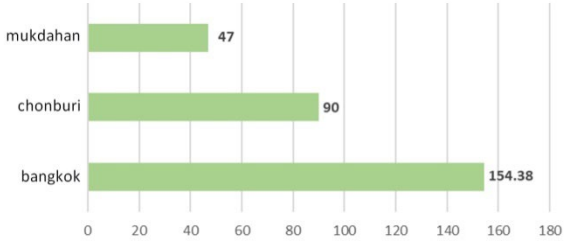


Fig. 5: Top 3Chi-square values for Province

3.3. Data Splitting

The dataset was divided into a training set, which consisted of 80% of the data and was used to train the machine learning model, and a test set, which consisted of 20% of the data and was used to evaluate the performance of the model on unseen data.

3.4. Class Weighting

For imbalanced data, where the proportion of data in each class differed significantly, one approach to address this was by assigning different weights to each class during the training process. In this research, the weight for the fatalities class, which was the minority class, was set to seven times the weight of the injured class, which was the majority class. This helped to balance the contribution of each class during training.

3.5. Machine Learning Model Training

The research developed machine learning models using five algorithms: XGBoost, Random Forest, Bagging, Decision Tree, and Multilayer Perceptron. The models were trained on labeled data with specific configurations and parameter settings. Described below were the parameter settings for each model:

The **XGBoost** model was based on a GBTree with 1000 gradient boosted trees (the number of boosting rounds), a learning rate of 0.01, and a maximum tree depth of 9.

In training the **Random Forest** model, the process employed the Gini index for split evaluation, set the number of trees to 100, with no limit on tree depth, restricted the number of attributes considered using the square root function, and used bootstrap samples when building trees.

For the **Bagging** model, Decision Trees were used as the base model, comprising a total of 10 trees. The tree attributes were sampled without replacement. Bootstrap samples were drawn with replacement, with the number of samples matching that of the original dataset.

For the **Decision Tree** model, the training process utilized the Gini index for calculations, imposed no limit on tree depth, and assigned a weight to the Fatality data that was seven times greater than that of the Injured data.

For the **Multilayer Perceptron** model, two hidden layers were created, each containing 100 neurons. The activation function used was ReLU (Rectified Linear Unit), and the Adam optimization algorithm was employed. The adaptive learning rate started at 0.01. The model was trained for up to 500 epochs.

3.6. Model Performance Evaluation

The trained models were evaluated using appropriate evaluation metrics depending on the nature of the problem and the desired outcomes. For this research, the traffic accident dataset was imbalanced because the number of injuries was significantly higher than the number of fatalities. Consequently, the accuracy evaluation metric [14] could not effectively assess the model's performance. Accuracy as an evaluation metric had limitations that could lead to the neglect of the minority class data, as it did not consider the class distribution and did not equally emphasize both groups of data. These limitations could introduce bias in the evaluation. Therefore, other evaluation metrics needed to be considered, including:

Precision: The ratio of correctly classified positive examples (True Positive: TP) to the total number of examples classified as positive, whether correctly or incorrectly predicted (False Positive: FP) [15], as expressed in Equation 1:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{1}$$

Recall or Sensitivity: The ratio of correctly classified positive examples (True Positive: TP) to the total number of actual positive examples of the class under consideration, including both true positives and false negatives (False Negative: FN), as expressed in Equation 2:

$$\text{Recall} = \frac{TP}{TP+FN} \tag{2}$$

F1-Score: The harmonic mean of Precision and Recall, providing a single metric that balanced both Precision and Recall, as calculated using Equation 3:

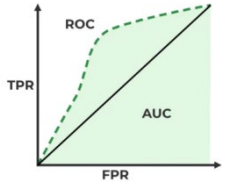
$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Balanced Accuracy: The average of the true positive rate (TPR) and the true negative rate (TNR) calculated for each class [16]. This metric measured the average accuracy obtained on both the minority class and the majority class, making it suitable for imbalanced datasets. It was calculated using Equation 4:

$$\text{Balanced Accuracy} = \frac{\text{TPR} + \text{TNR}}{2} \quad (4)$$

$$\text{where TPR} = \text{Recall} = \frac{TP}{TP + FN} \quad \text{and} \quad \text{TNR} = \frac{TN}{TN + FP}$$

The ROC Curve and Area Under the ROC Curve (ROC-AUC): The relationship between the true positive rate (TPR) and the false positive rate (FPR) [17], with the area under the ROC Curve (AUC) derived from the ROC Curve, indicating how well the model could distinguish between classes. The ROC Curve was plotted with TPR (y-axis) against FPR (x-axis), as shown in Fig. 6.



where $\text{FPR} = \frac{FP}{FP + TN}$

Fig. 6: ROC-AUC curve

4. Results and Discussion

From creating models to predict the severity of traffic accidents using various algorithms and applying class weighting techniques, where higher weights were assigned to the minority class during training to reduce the imbalance, the results demonstrated the performance of different models in predicting severity levels. Each model was evaluated using various metrics, including Precision, Recall, F1-Score, Balanced Accuracy, and ROC-AUC.

The following section presents a detailed comparison of the overall performance of each model. The summarized results are shown in Table 3 as follows:

Table 3: Comparison of algorithm performance

Model	Class 0 (Injured)			Class 1 (Fatal)			Balanced Accuracy	ROC AUC
	Precision	Recall	F1-Score	Precision	Recall	F1-Score		
XGBoost	**0.95**	**0.98**	**0.96**	**0.78**	**0.57**	**0.66**	**0.77**	**0.90**
Random Forest	0.93	0.99	0.96	0.89	0.46	0.61	0.73	0.89
Bagging	0.94	0.98	0.96	0.76	0.51	0.61	0.74	0.81
Decision Tree	0.95	0.93	0.94	0.52	0.61	0.56	0.76	0.77
MLP	0.94	0.94	0.94	0.55	0.56	0.56	0.75	0.83

For the analysis of the ROC-AUC curve in each model, the ROC-AUC curve provides deep insights into the model's ability to distinguish between the severity levels of fatalities and injuries. The ROC-AUC graph illustrates the trade-off between recall and specificity, as shown in the results presented in Fig. 7. This comprehensive model comparison aids in selecting the most efficient model for predicting the severity level of accidents.

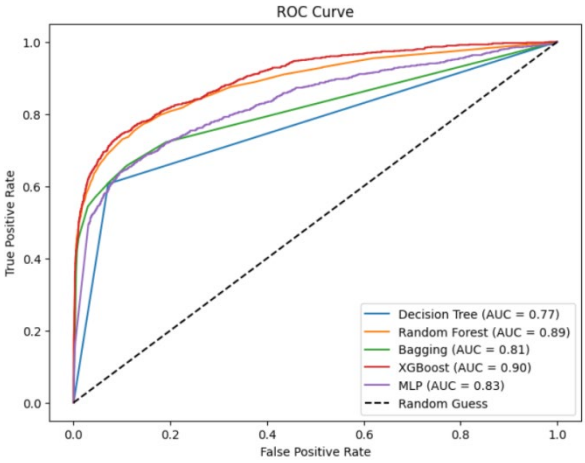


Fig. 7: Comparison of ROC-AUC curves for each algorithm of class 1

For this research, a significant importance was placed on class 1 (accidents with a severity level of Fatal) due to its significant impact on individuals, families, and communities. The research found that the XGBoost model performed better overall than other models in predicting the accident severity levels, particularly in Fatal cases. The model achieved a precision of 78% in identifying all cases predicted as fatalities, a recall of 57% indicating the proportion of actual fatalities identified by the model, an F1-score of 66% which was a balanced score between precision and recall, a balanced accuracy of 77% indicating overall accuracy of the model in predicting both classes considering the balance of data in each class, and a ROC-AUC score of 90% demonstrating the model's strong ability to differentiate between fatalities and injuries.

As for other models, the performance numbers in Table 3 shows no clear indication of next best models. However, the ROC-AUC curve in Fig. 7 shows that Random Forest and MLP, respectively, had the next best performance.

The Chi-square test indicated that the top three predicting attributes were vehicle type (motorcycles), accident type (pedestrian collisions), and accident type (opposite direction collision not overtaking) as its top three predictors. XGBoost identified vehicle type (motorcycles), vehicle type (pedestrians), and accident type (pedestrian collisions) as its top three predictors. For the Random Forest model, the most important attributes were vehicle type (motorcycles), number of injuries, and number of vehicles involved. Bagging highlighted number of injuries, vehicle type (motorcycles), and vehicle type (pedestrians) as the leading predictors. Lastly, the Decision Tree model indicated vehicle type (motorcycles), vehicle type (pedestrians), and number of injuries as the top three predicting attributes.

In summary, most models agreed that the key attribute in predicting severity was vehicle type, particularly when motorcycles were involved. Other important attributes included pedestrian involvement, and the number of injuries.

5. Conclusion

This research aimed to develop a model to predict the severity level of road traffic accidents resulting in fatalities and injuries. The road traffic accident dataset included various attributes such as accident location, vehicle type, accident type, suspected cause, and others. The predictive models were created using five machine learning algorithms: XGBoost, Random Forest, Bagging, Decision Tree, and Multilayer Perceptron.

It was found that the XGBoost model outperformed the other models in predicting the severity of road traffic accidents. This superiority was evident in its Precision, Recall, F1-Score, Balanced Accuracy, and ROC-AUC curve analysis. The next best models were Random Forest and MLP, although the differences among all the models were not substantial. Some key attributes that played a significant role in determining the target variable were vehicle type (motorcycles), pedestrian involvement, and the number of injuries.

For future research involving similar datasets, the XGBoost model can be enhanced to achieve better overall performance, particularly in managing imbalanced datasets. Additionally, further analysis of the XGBoost model can offer deeper insights into the significance of various factors influencing the severity prediction.

Acknowledgements

ChatGPT (https://openai.com) assisted in translating parts of this paper from its original language to English. The entire content was initially written by the authors and subsequently translated by ChatGPT. No part of the paper was generated by ChatGPT without corresponding original content from the authors. The authors reviewed, revised, and added content after the translation.

References

[1] World Health Organization. (2022, Jun 24). Number of road traffic deaths [Online]. Available: https://www.who.int/data/gho/data/themes/road-safety
[2] Ministry of Transport. (2022, Jun 24). Accidents on Ministry of Transport's Road Network [Online]. Available: https://datagov.mot.go.th/dataset/roadaccident
[3] H. Hala, C. Anass, B. Rajaa, B. Youssef and J. Garza-Reyes, "Machine learning techniques for forecasting the traffic accident severity," in *Proceedings of the International Conference on Digital Age & Technological Advances for Sustainable Development (ICDATA)*, Marrakech, Morocco, 2021.
[4] S. Malik, H. E. Sayed, M. A. Khan and M. J. Khan, "Road Accident Severity Prediction — A Comparative Analysis of Machine Learning Algorithms," in *Proceedings of IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT)*, Dubai, United Arab Emirates, 2021, pp. 69-74.
[5] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, United States, 2016, pp. 785–794.
[6] C. Kittinarador. (2022, Jun 24). Random Forest. GitHub [Online]. Available: https://guopai.github.io/ml-blog10.html
[7] IBM. (2023, May 20). What is Bagging? [Online]. Available: https://www.ibm.com/topics/bagging
[8] R. Klatchuen and C. Saenrat, "An Efficiency Comparison of Algorithms and Feature Selection Methods for Predict the Learning Achievement of Vocational Students," *Rajamangala University of Technology Thanyaburi Research Journal*, vol. 17, no. 1, pp. 1-10, 2018.
[9] P. Shukla. (2023, Jul 10). Multi-Layer Perceptrons: Notations and Trainable Parameters [Online]. Available: https://www.analyticsvidhya.com/blog/2022/10/multi-layer-perceptrons-notations-and-trainable-parameters
[10] W. Buathong. (2022, Jun 24). Chapter 3 Data Preprocessing [Online]. Available: https://wipawanblog.files.wordpress.com/2014/06/chapter-3-data-preprocessing.pdf
[11] J. Brownlee. (2023, May 25). One-Hot Encode Data in Machine Learning? [Online]. Available: https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning
[12] A. Biswal. (2023, May 25). What is a Chi-Square Test? Formula, Examples & Application [Online]. Available: https://www.simplilearn.com/tutorials/statistics-tutorial/chi-square-test
[13] A. Kumar. (2023, May 25). Dealing with Class Imbalance in Python: Techniques [Online]. Available: https://vitalflux.com/class-imbalance-class-weight-python-sklearn
[14] A. Allwright. (2023, May 25). Metrics for imbalanced data [Online]. Available: https://stephenallwright.com/imbalanced-data-metric
[15] O. Shalev. (2023, May 25). Recall, Precision, F1, ROC, AUC, and everything [Online]. Available: https://medium.com/swlh/recall-precision-f1-roc-auc-and-everything-542aedf322b9
[16] J. Akosa. (2023, May 25). Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data [Online]. Available: https://support.sas.com/resources/papers/proceedings17/0942-2017.pdf
[17] A. Bhandari. (2023, May 25). Guide to AUC ROC Curve in Machine Learning: What Is Specificity? [Online]. Available: https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning